Survival Analysis

PROJECT 2 REPORT

Introduction

For the lung cancer dataset, I will analyze the survival data by age in years. More precisely, I will focus on two age groups for patients with advanced lung cancer. Group 1 is the group where patients are younger than 65 years, and group 2 is the group where patients are 65 years or older. The three variables of most importance for the data analysis include the following: time (represents survival time in days), status (1 denotes censoring time; 2 denotes failure time), and age.

Research Interests and Questions

For the survival analysis, there are certain interests and questions that are to be investigated.

I. Research interest: Estimate the survival functions of the two age groups.
    i. Question: What is the survival probability by age ("< 65" and " ≥ 65")?
        i. Estimate the survival function for group 1 where the advanced lung cancer patients are younger than 65 years.
        ii. Estimate the survival function for group 2 where the advanced lung cancer patients are 65 years or older.

II. Research interest: Compare the survival experience of the two age groups.
    i. Question: Is there a significant difference in survival between the two age groups ("< 65" and " ≥ 65")?

Methods

To perform the survival analysis, I am using R. To estimate the survival functions of the two age groups, I am using the Kaplan-Meier estimator; this allows me to take the censored data into account. To compare the survival experience of the two age groups, I am using the Logrank test; the test will be conducted at a 5% significance level. The relevant R code for the two methods (Kaplan-Meier estimator and Logrank test) is included below.

I. Research interest: Estimate the survival functions of the two age groups.
    i. Question: What is the survival probability by age ("< 65" and " ≥ 65")?

R code:

```
library(survival)
library(survminer)
data(lung)
# classify age into categories to treat it as a group indicator
lung$age <- ifelse(lung$age<65, '65-','65+')

# Research Interest 1
fit.lung <- survfit(Surv(time, status)~age, data=lung)
```

```
# estimate of survival functions by age group.
summary(fit.lung)
# KM survival curves for the two age groups
ggsurvplot(fit.lung, conf.int = T, pval = T, xlab = "Time in days", risk.table = T)
```

II. Research interest: Compare the survival experience of the two age groups.
   i. Question: Is there a significant difference in survival between the two age groups ("< 65" and " ≥ 65")?

$H_0: S_1(\cdot) = S_2(\cdot)\ versus\ H_1: S_1(\cdot) \neq S_2(\cdot)$

R code:

```
# Research Interest 2
# Logrank test
diff.fit.lung <- survdiff(Surv(time, status)~age, data=lung)
diff.fit.lung
```

Results and Conclusions

For the research interest and question in part I., here are the results using the method Kaplan-Meier estimator to estimate the survival functions for the two age groups. The KM survival estimates at ordered distinct failure times are located in the "survival" column of the output.

Call: survfit(formula = Surv(time, status) ~ age, data = lung)

age=65-

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
53	128	1	0.9922	0.00778	0.9771	1.000
60	127	1	0.9844	0.01096	0.9631	1.000
61	126	1	0.9766	0.01337	0.9507	1.000
65	125	1	0.9688	0.01538	0.9391	0.999
71	124	1	0.9609	0.01712	0.9280	0.995
79	123	1	0.9531	0.01868	0.9172	0.990
81	122	2	0.9375	0.02140	0.8965	0.980
88	120	1	0.9297	0.02260	0.8864	0.975
92	119	1	0.9219	0.02372	0.8765	0.970
95	117	1	0.9140	0.02479	0.8667	0.964
105	116	1	0.9061	0.02580	0.8569	0.958
107	115	2	0.8904	0.02765	0.8378	0.946
110	113	1	0.8825	0.02851	0.8283	0.940
122	112	1	0.8746	0.02932	0.8190	0.934
131	111	1	0.8667	0.03010	0.8097	0.928
132	110	2	0.8510	0.03155	0.7913	0.915

135	108	1	0.8431	0.03222	0.7822	0.909
142	107	1	0.8352	0.03287	0.7732	0.902
145	106	2	0.8194	0.03409	0.7553	0.889
147	104	1	0.8116	0.03466	0.7464	0.882
156	103	1	0.8037	0.03521	0.7376	0.876
163	102	1	0.7958	0.03573	0.7288	0.869
166	101	1	0.7879	0.03624	0.7200	0.862
167	100	1	0.7800	0.03672	0.7113	0.855
170	99	1	0.7722	0.03719	0.7026	0.849
177	96	1	0.7641	0.03766	0.6938	0.842
179	94	1	0.7560	0.03812	0.6848	0.835
180	93	1	0.7479	0.03857	0.6760	0.827
181	92	2	0.7316	0.03941	0.6583	0.813
182	90	1	0.7235	0.03980	0.6495	0.806
186	89	1	0.7154	0.04018	0.6408	0.799
189	88	1	0.7072	0.04053	0.6321	0.791
194	85	1	0.6989	0.04090	0.6232	0.784
197	83	1	0.6905	0.04127	0.6142	0.776
199	82	1	0.6821	0.04161	0.6052	0.769
202	81	1	0.6736	0.04194	0.5963	0.761
210	79	1	0.6651	0.04227	0.5872	0.753
212	78	1	0.6566	0.04258	0.5782	0.746
218	77	1	0.6481	0.04287	0.5693	0.738
223	76	1	0.6395	0.04315	0.5603	0.730
226	73	1	0.6308	0.04344	0.5511	0.722
239	71	1	0.6219	0.04372	0.5418	0.714
245	68	1	0.6127	0.04403	0.5323	0.705
246	67	1	0.6036	0.04431	0.5227	0.697
268	63	1	0.5940	0.04463	0.5127	0.688
286	59	1	0.5839	0.04499	0.5021	0.679
291	58	1	0.5739	0.04533	0.4916	0.670
293	56	1	0.5636	0.04567	0.4809	0.661
305	51	1	0.5526	0.04609	0.4692	0.651
320	49	1	0.5413	0.04651	0.4574	0.641
337	47	1	0.5298	0.04692	0.4454	0.630
340	46	1	0.5183	0.04729	0.4334	0.620
345	45	1	0.5068	0.04762	0.4215	0.609
348	44	1	0.4952	0.04791	0.4097	0.599
353	43	1	0.4837	0.04816	0.3980	0.588
363	41	1	0.4719	0.04841	0.3860	0.577
364	40	1	0.4601	0.04862	0.3741	0.566
371	39	2	0.4365	0.04891	0.3505	0.544
387	34	1	0.4237	0.04912	0.3376	0.532
390	33	1	0.4108	0.04928	0.3248	0.520

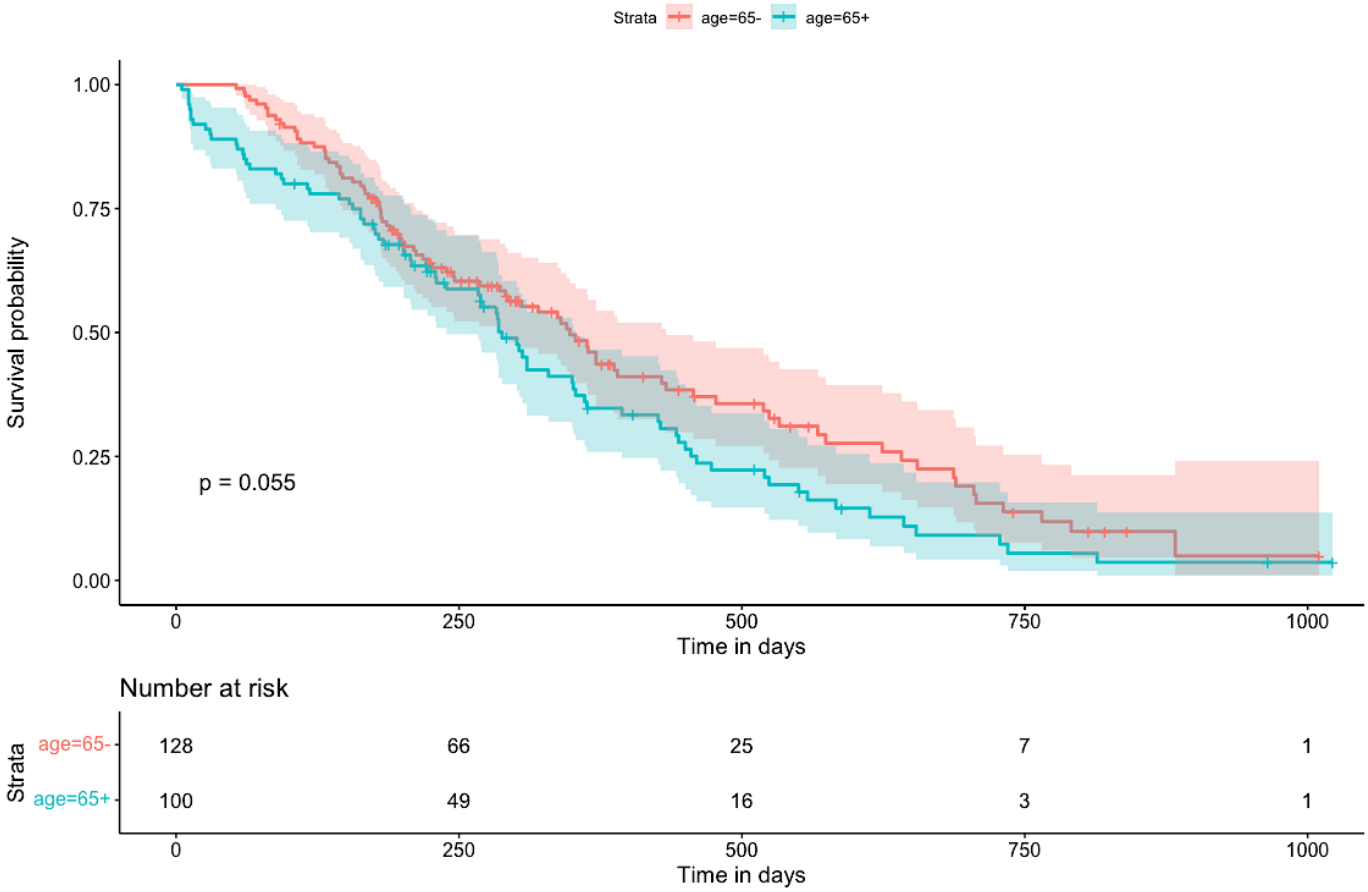
429	31	1	0.3976	0.04944	0.3116	0.507
433	30	1	0.3843	0.04954	0.2985	0.495
457	28	1	0.3706	0.04964	0.2851	0.482
477	26	1	0.3564	0.04973	0.2711	0.468
519	24	1	0.3415	0.04983	0.2566	0.455
524	23	1	0.3267	0.04982	0.2423	0.440
533	21	1	0.3111	0.04982	0.2273	0.426
567	18	1	0.2938	0.04996	0.2105	0.410
574	17	1	0.2765	0.04992	0.1941	0.394
624	16	1	0.2593	0.04970	0.1780	0.378
641	15	1	0.2420	0.04930	0.1623	0.361
655	14	1	0.2247	0.04872	0.1469	0.344
687	13	1	0.2074	0.04794	0.1319	0.326
689	12	1	0.1901	0.04696	0.1172	0.309
705	11	1	0.1728	0.04576	0.1029	0.290
707	10	1	0.1556	0.04433	0.0890	0.272
731	9	1	0.1383	0.04264	0.0756	0.253
765	7	1	0.1185	0.04087	0.0603	0.233
791	6	1	0.0988	0.03853	0.0460	0.212
883	2	1	0.0494	0.03988	0.0101	0.240

age=65+

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	100	1	0.9900	0.00995	0.97069	1.000
11	99	3	0.9600	0.01960	0.92235	0.999
12	96	1	0.9500	0.02179	0.90823	0.994
13	95	2	0.9300	0.02551	0.88131	0.981
15	93	1	0.9200	0.02713	0.86833	0.975
26	92	1	0.9100	0.02862	0.85560	0.968
30	91	1	0.9000	0.03000	0.84308	0.961
31	90	1	0.8900	0.03129	0.83074	0.953
53	89	1	0.8800	0.03250	0.81856	0.946
54	88	1	0.8700	0.03363	0.80652	0.938
59	87	1	0.8600	0.03470	0.79461	0.931
60	86	1	0.8500	0.03571	0.78282	0.923
62	85	1	0.8400	0.03666	0.77113	0.915
65	84	1	0.8300	0.03756	0.75955	0.907
88	83	1	0.8200	0.03842	0.74805	0.899
93	82	1	0.8100	0.03923	0.73665	0.891
95	81	1	0.8000	0.04000	0.72532	0.882
116	79	1	0.7899	0.04076	0.71390	0.874
118	78	1	0.7797	0.04147	0.70256	0.865
144	77	1	0.7696	0.04215	0.69128	0.857
153	76	1	0.7595	0.04280	0.68008	0.848

156	75	1	0.7494	0.04341	0.66894	0.839
163	74	2	0.7291	0.04453	0.64685	0.822
166	72	1	0.7190	0.04505	0.63589	0.813
175	70	1	0.7087	0.04556	0.62481	0.804
176	69	1	0.6984	0.04605	0.61378	0.795
179	68	1	0.6882	0.04650	0.60281	0.786
183	67	1	0.6779	0.04693	0.59189	0.776
201	63	2	0.6564	0.04784	0.56900	0.757
207	60	1	0.6454	0.04828	0.55743	0.747
208	59	1	0.6345	0.04868	0.54591	0.737
222	56	1	0.6232	0.04912	0.53397	0.727
229	53	1	0.6114	0.04958	0.52157	0.717
230	52	1	0.5997	0.05000	0.50925	0.706
239	50	1	0.5877	0.05042	0.49671	0.695
267	49	1	0.5757	0.05079	0.48425	0.684
269	48	1	0.5637	0.05113	0.47186	0.673
270	46	1	0.5514	0.05147	0.45924	0.662
283	44	1	0.5389	0.05180	0.44635	0.651
284	43	1	0.5264	0.05209	0.43355	0.639
285	42	2	0.5013	0.05254	0.40821	0.616
288	40	1	0.4888	0.05270	0.39566	0.604
301	38	1	0.4759	0.05286	0.38280	0.592
303	37	1	0.4630	0.05297	0.37003	0.579
306	36	1	0.4502	0.05304	0.35735	0.567
310	35	2	0.4245	0.05304	0.33225	0.542
329	33	1	0.4116	0.05296	0.31984	0.530
350	32	1	0.3987	0.05285	0.30751	0.517
351	31	1	0.3859	0.05269	0.29527	0.504
353	30	1	0.3730	0.05248	0.28311	0.491
361	29	1	0.3601	0.05222	0.27105	0.479
363	28	1	0.3473	0.05191	0.25908	0.466
394	26	1	0.3339	0.05161	0.24666	0.452
426	24	1	0.3200	0.05130	0.23373	0.438
428	23	1	0.3061	0.05092	0.22093	0.424
442	22	1	0.2922	0.05047	0.20827	0.410
444	21	1	0.2783	0.04995	0.19574	0.396
450	20	1	0.2644	0.04935	0.18335	0.381
455	19	1	0.2504	0.04867	0.17111	0.367
460	18	1	0.2365	0.04792	0.15901	0.352
473	17	1	0.2226	0.04708	0.14708	0.337
520	15	1	0.2078	0.04622	0.13435	0.321
524	14	1	0.1929	0.04524	0.12185	0.305
550	13	1	0.1781	0.04412	0.10958	0.289
558	11	1	0.1619	0.04298	0.09622	0.272

583	10	1	0.1457	0.04162	0.08325	0.255
613	8	1	0.1275	0.04021	0.06872	0.237
643	7	1	0.1093	0.03837	0.05492	0.217
654	6	1	0.0911	0.03604	0.04193	0.198
728	5	1	0.0729	0.03311	0.02989	0.178
735	4	1	0.0546	0.02942	0.01902	0.157
814	3	1	0.0364	0.02461	0.00969	0.137



We computed the survival probability by age ("< 65" and " $\geq$ 65") and plotted the KM survival probability against time in days. Based on the plot, there does not seem to be much of a difference (visually) between the two survival curves. To determine if the two age groups differ significantly in survival, I will use the Logrank test.

For the research interest and question in part II., here are the results using the Logrank test to compare the survival functions of the two age groups to determine if they differ significantly from one another.

Call:
survdiff(formula = Surv(time, status) ~ age, data = lung)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
age=65-	128	86	98	1.48	3.67
age=65+	100	79	67	2.17	3.67

Chisq= 3.7 on 1 degrees of freedom, **p= 0.06**

The p-value is 0.06 for the Logrank test. Since p-value=0.06 > α=0.05, we fail to reject H_0. There is no significant difference in survival between the two age groups ("< 65" and " $\geq$ 65").

References

[1] http://www.sthda.com/english/wiki/survival-analysis-basics